Инженерный вестник Дона, №4 (2016)
ivdon.ru/ru/magazine/archive/n4y2016/3794

Тематическая модель рейтингования интернет-сайтов по критерию социальной значимости

А.В. Сироткин, С.А. Шарыпов

Северо-Восточный государственный университет, Магадан

Аннотация: Рассматривается модель численной оценки информационного содержимого Интернет-сайтов, основанная на тематическом анализе текстов. Оценка производится с использованием частных качественных критериев, соответствующих принятым «ценностям общества». Численное значение соответствия получается с использованием статистических методов тематического анализа. Полученный результат используется для получения обобщённого рейтинга сайта, устанавливающего его социальную значимость.
Ключевые слова: тематический анализ, рейтинг сайта, ценность общества, ранжирование, лингвистический ключ.

Существует проблема неконтролируемого влияния информации, размещённой в Интернете (сайтов) на сознание неподготовленного пользователя. Это представляет определённую опасность, последствия которой отражаются в правовой, социальной, психологической, педагогической и других сферах деятельности и существования человека [1]. Предпринимаемые государством меры по ограничению доступа к таким ресурсам не исчерпывают всех возможных способов предупреждения этой опасности, в силу чего заинтересованными разработчиками создаются различные программные средства информирования пользователя или ограждения его от негативного воздействия опасных ресурсов. Одним из таких средств может выступать система информирования пользователя о соответствии содержимого сайта принятым ценностями общества или критериям желательности информационного контента на основе проставления рейтинга, в цифровом виде отражающего это соответствие.

Данную задачу можно разбить на две подзадачи, каждая из которых имеет право на самостоятельное решение.

Первая из них – это ранжирование текста опубликованного в Интернете документа по тематической направленности с учётом сформулированного множества тематических критериев.

Вторая – построение обобщённого показателя, свёртывающего частные тематические показатели с целью получения численной оценки соответствия некоему абстрактному критерию, сформированному на множестве частных тематических критериев с учётом множества сформулированных критериев качества. В данной работе основное внимание уделяется решению первой задачи путём формализации её основных исходных условий и используемых методов решения.

Информационные объекты, размещённые в Интернете, можно различать, прежде всего, по тематике содержания, используемой лексике, а также в прагматическом смысле: по желательности для пользовательской аудитории и соответствии принятым ценностям общества. Существует множество подходов к ранжированию информационных Интернет-ресурсов, основанных, например, на оценках пользователей (*WOT*, *http://www.mywot.com*), или, например, с точки зрения поисковой системы [2-3]. Многообразие способов оценки на этом перечне не исчерпывается, оставляя возможности для разработки иных методик, в числе которых может быть использована численная оценка соответствия материала сайта принятым качественных критериям, основанным на содержании размещённого документа.

Исследования, проведённые в области лингвистического анализа содержимого Интернет-ресурсов, в том числе и с участием авторов [4-6], позволили сформулировать множества критериев положительного (позитивного) и отрицательного (негативного) характера, отражающих соответствие сайтов составу ценностей общества, а также на основе социологических опросов определить критерии тематической желательности

текстов и установить весовые параметры для каждого из них. Данные множества были определены как множество «ценностей общества» и множество критериев «желательности контента». Эти множества могут быть использованы в качестве критериальной базы для построения автоматизированной системы рейтингования информационных Интернет-ресурсов как по критериям желательного контента, так и в соответствии с составом «ценностей общества».

Поскольку и в том и в другом случае речь идёт о содержании текстов информационных объектов, то в рамках поставленной задачи рейтингования следует, прежде всего, исследовать применимость методов тематического анализа для многокритериальной оценки документов. В настоящее время достаточно большое внимание уделяется вероятностным методам, основанным на статистике встречаемости синтаксических структур в тексте документа, особенно при использовании средств автоматизации для его обработки. Для решения поставленной задачи представляется возможным построение автоматизированной системы рейтингования сайтов путём свёртки частных показателей, полученных на основе упрощённой модели тематического анализа содержимого Интернет-ресурсов.

Известны и широко используются, например, для тематической классификации документов, различные методы, в том числе векторная модель (*Vector Space Model, VSM*) [7], предназначенная для назначения весов методом *TF-IDF*, которая в общем виде применительно к одному документу, формулируется как:

$$TF(t,d) = \frac{freq(t,d)}{\max_{w \in D} freq(w,d)}, \quad (1)$$

где *TF* (*term frequency*) — нормализованная частота слова в тексте, в знаменателе общее количество слов в тексте.

Известны модификации *TF-IDF*, например модель, учитывающая длину документа [8]:

$$TF' = \frac{TF}{TF + 0.5 + 1.5\dfrac{len_d}{len_{avg}}}, \qquad (2)$$

где len_d – длина документа, len_{avg} – средняя длина документа.

Представленная модель, основанная на статистическом тематическом анализе текста документа, представляется наиболее применимой для решения задачи рейтингования информационных Интернет-ресурсов. В обоснование этого утверждения можно привести следующие доводы, основанные на работах исследователей, например [9].

1. Предметность восприятия текста формируется на основе созданных тематических образов.
2. Признаки, идентифицирующие тематический образ, определяются на основе лингвистической экспертизы.
3. Распознавание текста осуществляется путём анализа лингвистических единиц, имеющего избирательный эвристический характер.

Образное тематическое восприятие текста документа Интернет-источника можно представить в виде графа, представляющего модифицированную репликацию графа [9], который приведён на рис. 1.

В приведённом графе экземпляры лексических единиц представлены словами, словосочетаниями и пр., составляющими текст Интернет-документа. Образы значимых тематических языковых единиц составляют множество тематических информационных объектов, составляющий общий информационный объект (документ), определённых для тематического анализа документа, и проецирующихся на множество «ценностей общества» или множество критериев «желательности контента». В нашей задаче каждый тематический образ, определяемый набором лексем, соответствует

одному или нескольким частным показателям, входящим в множества «ценностей общества» или «желательности контента».

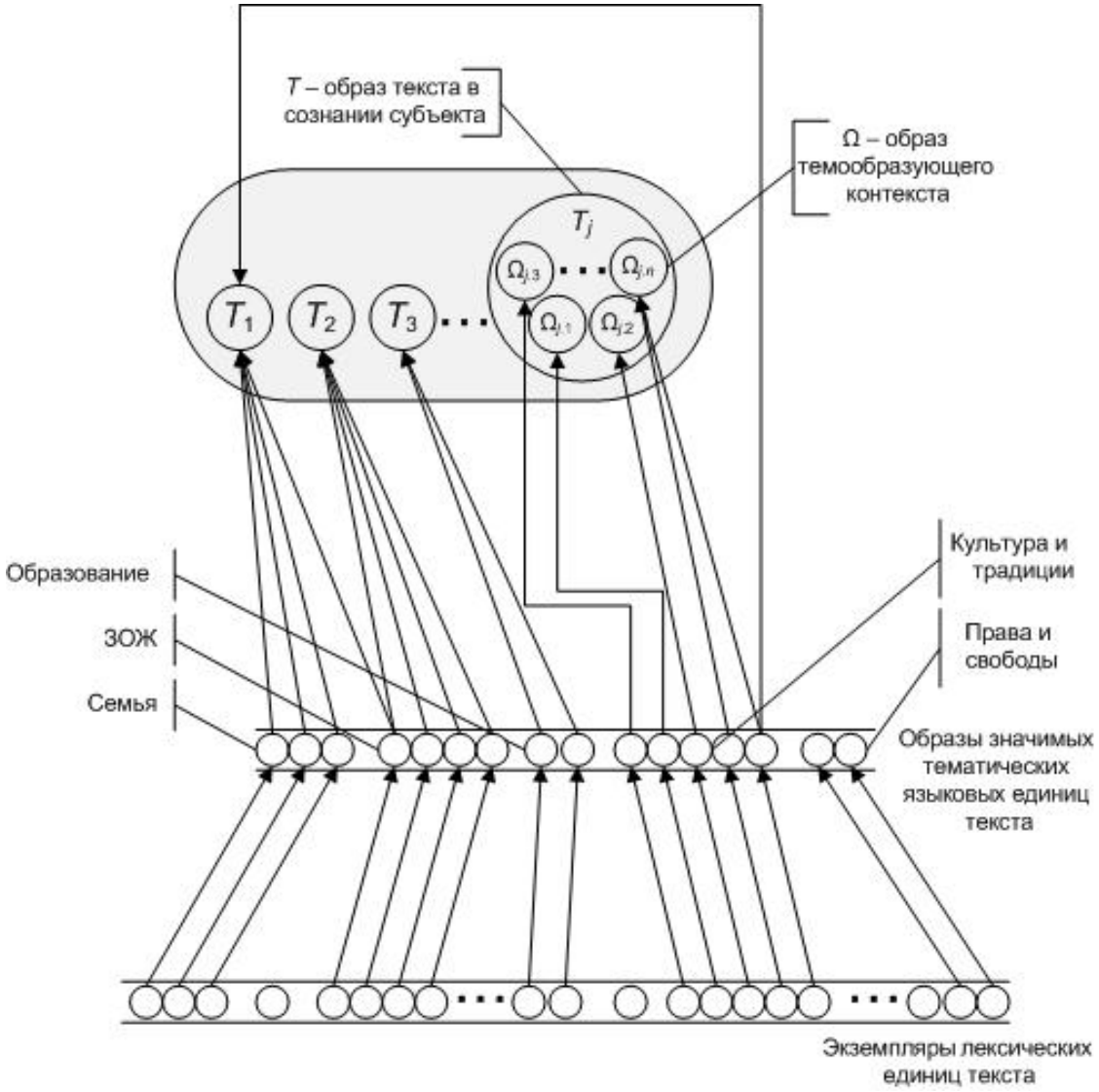

Рис. 1. – Образная интерпретация тематического восприятия содержимого Интернет-страницы

На этапе реализации необходимо было сформулировать взаимосвязи между множествами показателей и анализируемыми информационными объектами. Анализ производился на основе лингвистических ключей, отражающих семантическое содержание объекта (информационного контента). Анализ проводился по нескольким направлениям:

1. тематическое содержание текста;

2. авторское отношение к описываемой теме;
3. использование ненормативной лексики.

Введём описание параметров и компонентов модели тематического анализа.

Пусть множество сайтов U, подлежащих оценке, которое содержит в себе *URL*-адреса отдельных i-х объектов $u_i \in U; i \in [\overline{1,I}]; I = |U|$.

Каждый сайт представляет собой i-й объект контента, состоящий из лингвистических объектов, подлежащих анализу. Из каждого объекта u_i может быть выделено соответствующее множество O_i, включающее в себя лингвистические объекты $o_{j.i} \in O_i; j \in [\overline{1,J}]; J = |O_i|$. Лингвистические объекты выделяются из текста контента, прошедшего очистку от *HTML*-тегов, включений фрагментов программных кодов *Java-script* или др., стилей *CSS* и пр.

Для лингвистического анализа используется множество словарей Q, каждый элемент которого – представляет собой словарь $q_l \in Q; l \in [\overline{1,L}]; L = |Q|$, семантически соответствующий определённому критерию. Все критерии составляют множество C, каждый элемент которого – c_l принимается как частный показатель оценки содержания по словарю.

Каждый q_l словарь содержит множество K_l лингвистических ключей $k_{m.l} \in K_l; m \in [\overline{1,M}]; M = |K_l|$. Каждый ключ является экземпляром словаря, содержащим атрибуты ключа, такие как содержание – $s_{m.l}$ и вес – $w_{m.l}$ в рамках действия критерия c_l. Можно сказать, что каждый ключ представляет собой кортеж $k_{m.l} = (s_{m.l}, w_{m.l})$.

Отношение между множествами Q и K можно описать как $q_i = K_l, K_l \subset Q$.

Отношения между элементом u_i и множеством O_i можно описать как соответствие, т.е. $u_i \to O_i$.

Задачей исследования является поиск множества A_i вхождений лингвистических ключей в объект $O_i \to u_i$, образуемого как пересечение $A_i = O_i \cap Q$ с последующим расчётом рейтинга $\Psi_i = F(A_i)$, где F – некий функционал, применяемый к множеству вхождений. Можно сформировать кортеж R_i значений частных показателей $r_{l.i} \in R_i$, соответствующих множеству критериев C как $h : C \to R_i$, каждый из которых определяется как

$$r_{l.i} = \sum_{m=1}^{M} \rho_{i.m.l} w_{m.l}, \quad (3)$$

где $\rho_{i.m.l}$ – количество вхождений $k_{m.l}$-го ключа в o_i-й объект. С учётом модели *TF-IDF* (1), учитывающий плотность l-х лингвистических ключей в тексте, выражение (3) будет выглядеть как

$$r_{l.i} = \frac{\sum_{m=1}^{M} \rho_{i.m.l} w_{m.l}}{W_i}, \quad (4)$$

где W_i – общее количество слов в i-м объекте. Выражение (4) более адекватно отражает тематическую направленность документа, поскольку не зависит от длины текста, которая может «размывать» тематический образ.

Обобщённый рейтинг i-го сайта, исходя из наличия рассчитанного кортежа значений частных показателей

$$r_i = (r_{1.i}, r_{2.i} \ldots, r_{l.i}), \quad (5)$$

рассчитывается как

$$R_i = f(G, r_i), \quad (6)$$

где f – функция свёртки (обобщающая функция), применяемая к набору показателей в условиях многокритериальной оценки.

Показатели (5) отражают качественные характеристики сайта в разной степени устанавливающие его соответствие ценностям общества. Это проявляется в множестве весовых коэффициентов

$$G = \{g_1, g_2, \cdots, g_l\}; \quad h: G \to R_i; \quad i \in [\overline{1, |O|}]. \qquad (7)$$

Исходя из (7) к формированию функции свёртки (6) можно применить различные методы из задач многокритериального анализа, включая лексимакс [10] или метод последовательных уступок, или др.

В общем, выражения (4-7) можно рассматривать как модель, описывающую поиск решения в задаче рейтингования сайтов по установленным тематических критериям. Вид обобщающей функции (6) в настоящее время не определён и является объектом исследования на дальнейших этапах работы. Полученная формальная модель может быть использована для решения задач тематического анализа и ранжирования текстов по различным критериям с использованием конечного множества лингвистических ключей.

Литература

1. Сироткин А.В., Брачун Т.А. Безопасность человека в Интернете. Магадан: Ноосфера, 2014. 186 с.

2. Пестерев П.В., Янишевская А.Г. Модель суммарной оценки сайта в сети Интернет на основе факторов ранжирования // Инженерный вестник Дона, 2015, №3 URL: ivdon.ru/ru/magazine/archive/n3y2015/3216.

3. Пестерев П.В., Дьяконов Д.В., Рудюк А.П., Янишевская А.Г. Влияние факторов ранжирования на позиции сайтов в поисковых системах // Инженерный вестник Дона, 2014, №4 URL: ivdon.ru/ru/magazine/archive/n4y2014/2729.

4. Барели Д.Г., Исмаилов Н.Р., Корниенко М.В., Протопопов А.С., Сироткин А.В. Анализ информационных предпочтений молодёжи в сети Интернет. // Северо-Восточный научный журнал. 2013, № 1. С. 13-17.

5. Шарыпов С. А. Автоматизация контентного рейтингования интернет-сайтов на основе лингвистического анализа // Научное сообщество студентов

XXI столетия. ТЕХНИЧЕСКИЕ НАУКИ: сб. ст. по мат. XXXI междунар. студ. науч.-практ. конф. № 4(30). URL: sibac.info/archive/technic/4(30).pdf (дата обращения: 01.11.2016).

6. Протопопов А.С., Сироткин А.В. Техническое решение защиты детей от интернет-угроз в Магадане. Концептуальное обоснование // Информационные технологии в обществе, образовании и науке. Материалы Международной научно-практической интернет-конференции 26-27 ноября 2013 г. / ответ. ред. Т.А. Брачун. Магадан: СВГУ. 2014. С. 167-175.

7. James Allan: James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. Topic Detection and Tracking Pilot Study. Final Report. Proceedings of the Broadcast News Transcription and Understanding Workshop (Sponsored by DARPA), Feb. 1998. 25 p.

8. Allan, J. and Lavrenko: Allan, J. and Lavrenko, V. and Malin, D. and Swan, R. Detections, bounds, and timelines: UMass and TDT-3. In Proceedings of Topic Detection and Tracking Workshop, pp. 167-174, Vienna, VA, 2000.

9. Мячина Е.В. Автоматизированный анализ текста на основе вероятностно-статистической модели и его применение в региональном законотворчестве: дис. ... канд. тех. наук: 05.25.05. М., 2002. 188 с.

10. Ozyurt S., Sanver R. A general impossibility result on strategy-proof social choice hyperfunctions. Games and Economic Behavior. 2009. Vol. 66. pp. 880-892.

References

1. Sirotkin A.V., Brachun T.A. Bezopasnost' cheloveka v Internete [Safety of the person on the Internet]. Magadan: Noosfera, 2014. 186 p.

2. Pesterev P.V., Janishevskaja A.G. Inženernyj vestnik Dona (Rus), 2015, №3 URL: ivdon.ru/ru/magazine/archive/n3y2015/3216.

3. Pesterev P.V., D'jakonov D.V., Rudjuk A.P., Janishevskaja A.G. Inženernyj vestnik Dona (Rus), 2014, №4 URL: ivdon.ru/ru/magazine/archive/n4y2014/2729.

4. Bareli D.G., Ismailov N.R., Kornienko M.V., Protopopov A.S., Sirotkin A.V. Severo-Vostochnyj nauchnyj zhurnal. 2013. № 1. pp. 13-17.

5. Sharypov S. A. Nauchnoe soobshhestvo studentov XXI stoletija. TEHNIChESKIE NAUKI: sb. st. po mat. XXXI mezhdunar. stud. nauch.-prakt. konf. (Proc. The International Student's Scientifically-Practical Internet Conference). № 4(30). URL: http://sibac.info/archive/technic/4(30).pdf

6. Protopopov A.S., Sirotkin A.V. Informacionnye tehnologii v obshhestve, obrazovanii i nauke. Materialy Mezhdunarodnoj nauchno-prakticheskoj internet-konferencii (Proc. The International Scientifically Practical Internet Conference). Magadan: SVGU, 2014, pp. 167-175.

7. James Allan: James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. Topic Detection and Tracking Pilot Study. Final Report. Proceedings of the Broadcast News Transcription and Understanding Workshop (Sponsored by DARPA), Feb. 1998. 25 p.

8. Allan, J. and Lavrenko: Allan, J. and Lavrenko, V. and Malin, D. and Swan, R. Detections, bounds, and timelines: UMass and TDT-3. In Proceedings of Topic Detection and Tracking Workshop, pp. 167-174, Vienna, VA, 2000.

9. Mjachina E.V. Avtomatizirovannyj analiz teksta na osnove verojatnostno-statisticheskoj modeli i ego primenenie v regional'nom zakonotvorchestve [The automated analysis of the text on the basis of is likelihood-statistical model and its application in regional lawmaking]: dis. ... kand. teh. nauk: 05.25.05. M., 2002. 188 p.

10. Ozyurt S., Sanver R. A general impossibility result on strategy-proof social choice hyperfunctions. Games and Economic Behavior. 2009. Vol. 66. pp. 880-892.